THE ASTROPHYSICAL JOURNAL, 564:495–507, 2002 January 1
© 2002. The American Astronomical Society. All rights reserved. Printed in U.S.A.

SOME TESTS TO ESTABLISH CONFIDENCE IN PLANETS DISCOVERED BY TRANSIT PHOTOMETRY

JON M. JENKINS
SETI Institute/NASA Ames Research Center, MS 245-3, Moffett Field, CA 94035; jjenkins@mail.arc.nasa.gov

AND

DOUGLAS A. CALDWELL[1] AND WILLIAM J. BORUCKI
NASA Ames Research Center, MS 245-3, Moffett Field, CA 94035; dcaldwell@mail.arc.nasa.gov, wborucki@mail.arc.nasa.gov
Received 2001 April 11; accepted 2001 August 30

ABSTRACT

Increased attention is being paid to transit photometry as a viable method for discovering or confirming detections of extrasolar planets. Several ground-based efforts are underway that target short-period, giant planets such as 51 Peg b, and several missions have been proposed to NASA and ESA to detect planets as small as Earth from spaceborne photometers. The success of these efforts depends in part on the ability to establish appropriate detection thresholds to control false alarm rates and the ability to assess the statistical confidence in planetary candidates drawn from any such search. This latter function attains higher importance for the space-based efforts, where direct ground-based confirmation of terrestrial-size planets is not possible. These tasks are complicated by the need to survey tens of thousands of stars to overcome the limited geometric probability of transit alignment and by the nature of the transit signals themselves. In this paper, we present empirical methods for setting appropriate detection thresholds and for establishing the confidence level in planetary candidates obtained from transit photometry of even a large number of stars. The methods are simple and allow the observer to quickly assess the statistical significance of any particular set of transits.

Subject headings: planetary systems — techniques: photometric

1. INTRODUCTION

With the astonishing discovery of about a dozen giant short-period (<7 day) planets in the last 5 yr, astronomers are turning to transit photometry to discover new planets and to confirm radial velocity detections (Borucki et al. 2001; Brown & Charbonneau 2000). In transit photometry, nearly continuous flux measurements of many individual stars are used to search for signatures caused by the transit of a planet crossing a stellar disk. The amplitude of the flux reduction reveals the size ratio of the planet to the star, whereas the time interval between transits is simply the orbital period. From Kepler's third law and knowledge of the stellar mass and size, the planetary size and semimajor axis can be determined (Borucki & Summers 1984; Schneider & Chevreton 1990). Recently, transit photometry confirmed the planetary nature of HD 209458b first by ground-based photometry (Charbonneau et al. 2000; Henry et al. 2000) and subsequently by space-based photometry (Castellano et al. 2000; Robichon & Arenou 2000). Transit photometry searches are not new. The Transits of Extrasolar Planets (TEP) Network (Deeg et al. 1998; Doyle et al. 2000; Jenkins, Doyle, & Deeg 2000) has been observing one of the smallest known eclipsing binary system, CM Draconis, for evidence of small transiting planets and large nontransiting ones since 1994. At least four groups are attempting to detect 51 Pegasi–type planets from the ground (see, e.g., Borucki et al. 2001; Brown & Charbonneau 2000; Henry et al. 2000; Howell et al. 2000), while others are using spaceborne instruments including the *Hubble Space Telescope* (Gilliland et al. 2000; Brown et al. 2001) and the *Hipparcos* data archive (Laughlin 2000). This paper focuses on the question of assessing the significance of any purported transit detections given the unique character of this detection technique as opposed to radial velocity and astrometric searches.

The radial velocity and astrometric methods of finding planets seek to detect harmonic signatures of the gravitational effect of a planet on its parent star. The former technique is sensitive to the changing projected velocity of the star along the line of sight from Earth. The latter method detects the presence of a planet from periodic changes in the apparent position of the parent star relative to nearby reference stars as it orbits the planet-star system barycenter. The signal detection problems for both of these techniques can be cast into the frequency domain, where a periodogram is formed and large positive spikes indicate the presence of a possible planet. Although eccentricity precludes such signals from being purely sinusoidal, the signatures always consist of an impulse train in the frequency domain with most of the energy at the fundamental frequency equal to the inverse orbital period. Black & Scargle (1982) examine the case for astrometry, while Cumming et al. (1999) examine the detection strategy and characterize the null statistics for an 11 yr long radial velocity survey. Once a planet is detected, the orbital parameters are varied until the predicted astrometry or radial velocities fit the observations in a least-squares sense.

The signature of a transiting planet in a single-star system, however, is fundamentally different from these two cases in that the signal consists of equally spaced pulses of duration much shorter than the orbital period. For example, HD 209458b exhibits $\sim 1.6\%$ deep transits lasting about 3 hr spaced 3.52 days apart (Brown & Charbonneau 2000). The Fourier transform of such a pulse train consists of an impulse train spaced at $T_p^{-1} = 0.2837\ \mathrm{day}^{-1}$ with an

[1] National Research Council Associate.

envelope tracing out the Fourier transform of an individual pulse, i.e., a function similar to a periodic $\mathrm{sinc}^2(\omega)$ function. As a consequence of its concentration in time, the Fourier transform (FT) of a transit pulse train is broadband in the frequency domain. In addition, observation noise accompanies any data set and provides continuous noise power and will thus dominate the power spectrum of the data, masking the spectral features of the transits except for very high signal-to-noise ratio (S/N) cases. Therefore it is difficult to detect transits directly in the frequency domain as the energy is not highly concentrated in a few periodogram bins. Transits are most easily detected in the time domain or in another domain that reflects the time-confined nature of the signal, such as an appropriate wavelet domain. The fact that the signal is periodic coupled with the unavailability of an efficient detection possibility in the frequency domain complicates the detection of such signals, as the period is not known a priori. (HD 209458b is an exception in that the planet was first detected through its radial velocity perturbations.)

For transit detection there are two basic parameters: period and phase (epoch of first transit). Transit duration and the detailed shapes of individual transit pulses are less important, as discussed in Jenkins, Doyle, & Cullers (1996). For a star with known mass and size, the maximum transit duration is constrained by the orbital period and the size of the parent star (Koch et al. 1998). The detection of planets via transit photometry requires a thorough search through this parameter space, by varying the period and phase (and to some degree transit duration) and thresholding a detection statistic obtained at each trial point in the parameter space. A detection statistic is simply a scalar obtained by a calculation performed on the data to test for the presence of a signal. Its value then can be interpreted in terms of its statistical significance, i.e., how likely it is that noise fluctuations alone caused the detection statistic to attain its observed value.

An appropriate criterion for establishing a detection threshold in this case is the Neyman-Pearson criterion: maximize the detection rate while achieving the given desired false alarm rate. For simple detection problems such as detecting the presence of a signal in each of a consecutive sequence of disjoint time intervals, all that is required is a knowledge of the statistical distribution of the detection statistics when no signal is present, the so-called null statistics. In our case, however, the detection statistics are not independent; there is a complex web of correlations among the large number of target signals. Thus, we need to understand the behavior of the maximum expected null statistic for a single light curve over all the possible planets we search for. There appears to be no closed-form description for the desired distribution even under the simplest of assumptions. This problem has been largely overlooked in the literature to date.

Several papers have appeared recently regarding transit photometry. Deeg, Favata, & the Eddington Science Team (1998) estimate the performance of the proposed ESA *Eddington* mission by conducting Monte Carlo simulations. Transits were added to white Gaussian noise sequences representing observational noise, and the ability of a matched filter algorithm to detect the transits was examined. The detection threshold was determined by examining the results of 10^5–10^7 trials without adding transits to the observational noise. The actual value of the threshold, however, is not given, and no further details are presented on the search space explored for each trial. This approach is similar to the methods presented here, but the focus was on the expected performance of *Eddington*. Our intent is to provide a general theoretical framework and an associated empirical methodology that can be applied to any transit photometry campaign. Two other papers, Gilliland et al. (2000) and Brown et al. (2001), contain discussions of the puzzling lack of planets in a survey of 35,000 main-sequence stars in the globular cluster 47 Tucanae. A detection threshold of 6.3 σ was used based on the assumption that the observation noise was white and Gaussian and that $\sim 6 \times 10^5$ actual tests were conducted per star. This is most likely a conservative approach given that the set of tests were highly correlated. A particularly compelling argument is presented demonstrating that the sensitivity required to detect 51 Pegasi–like planets in 47 Tuc was achieved.

Under certain conditions, the detection threshold can be determined according to a heuristic method developed by K. Cullers (1995, private communication). The number of effective independent tests for a well-sampled, contiguous data set can be estimated in the following way: take the number of independent phases for the longest orbital period sought, square it, and multiply the result by the number of stars to be surveyed. That is, the first transit must occur sometime during the first section of data that is as long as the longest period under consideration, and the last transit must occur somewhere in the last section of data that is as long as the longest orbital period under consideration. The number of effective independent tests is equal to the number of permutations of possible independent phases for the first transit and those for the last transit. While this approach can be proved to yield the correct answer, it is not clear how to apply it to all cases of interest. In particular, if the time series obtained are not well sampled or contiguous, or the length of the data set is only as long as the longest period of interest, the assumptions underlying this method do not apply. Moreover, this method makes the following assumptions: (1) Consider each test as a set of indices specifying which points are "in transit." The set obtained by taking the "exclusive or" operation of any two tests' "transit point sets" is at least as large as the number of points in a single transit event. (2) For any given test there is always another test where the first constraint is strictly met; i.e., the nonoverlap between the two tests is exactly one transit duration. If the grid of the search space is more sparse or more dense, this method might over- or underestimate the effective number of independent statistical tests conducted in the search. In any event, a detailed understanding of the null statistics for transit campaigns is crucial to the success of both ongoing and proposed missions and projects.

In this paper, we show that the problem of setting the required detection threshold can be broken into two parts: (1) the establishment of the effective number of independent statistical tests conducted in searching a single light curve for planetary transits, N_{EIT}, and (2) the characterization of the null statistics for a light curve with a candidate sequence of transits. The latter determines the false alarm rate for a single test for that particular star as a function of the detection threshold. The former quantity, together with the total number of target stars in the program, dictates the requisite single-test false alarm rate for a given desired number of false alarms for the entire experiment. These two quantities are the subject of this paper, in which we seek to present

empirical methods for estimating them. Moreover, the methods presented take the characteristics of the noise distribution and its temporal correlation structure directly into account. The results, then, provide robust tools for evaluating transit photometry campaigns and for establishing confidence intervals on planetary candidates.

We describe one method for determining the equivalent number of independent statistical tests conducted in searching over a restricted region in parameter space for transits in a photometric data set in § 2. We present two different bootstrap methods for establishing the confidence in a planetary candidate from a given light curve in § 3 and in § 4. The first of these is appropriate for data sets like the *Hipparcos* archive, where sampling is incredibly sparse (89 points spread over 3 years' time in the case of HD 209458). The second is tailored for data sets similar to those obtained for the NASA Ames Vulcan Camera Project where 6–10 hr of data at more than 4 hr^{-1} is collected every possible night for up to 12 weeks each year.

The results of this paper show that transit photometry is a promising method for detecting planets even in the presence of non-Gaussian, colored noise and with the required large number of target stars ($>100{,}000$ stars in the case of the *Kepler* mission; Borucki et al. 1997) for the small geometric probability of transit alignment.

2. THE EQUIVALENT NUMBER OF STATISTICAL TESTS CONDUCTED IN SEARCHING A LIGHT CURVE FOR PLANETARY TRANSITS

In this section we discuss the problem of determining the equivalent number of statistical tests conducted in searching a photometric data set for transiting planets. A similar problem is encountered in detecting sinusoidal signals in noise-corrupted time series. Horne & Baliunas (1986) proposed a Monte Carlo technique for determining the effective number of independent frequency bins in the Lomb-Scargle periodogram of a time series, an essential step in determining an appropriate detection threshold and for assessing the statistical significance of any peak in the periodogram. Here we propose an analogous approach for the transit detection problem. To open the discussion, we review some basic detection theory relevant to the problem and then illustrate various facets of the problem for non-Gaussian noise. We provide an argument supporting the validity of the results derived for white Gaussian noise to more general cases of colored non-Gaussian noise. We proceed with the case of white Gaussian observation noise, giving a prescription for determining the effective number of independent tests. This is followed by several examples drawn from actual or anticipated observations.

If we wish to detect a deterministic signal in a noisy data set where the noise is Gaussian (colored or white), the optimal detector consists of a prewhitening filter followed by a matched filter detector (see Kay 1998). For the transit detection problem, a whitening filter can be thought of in terms of detrending the light curve to make it possible for a simple matched filter to detect a transit. Simple matched filters do *not* take into consideration points "out of transit." Thus, if the transits are superposed upon a slowly varying background with large excursions compared to the depth of transit, and if no prewhitening is performed, the matched filter will have a difficult time distinguishing transits from negative excursions occurring on longer timescales. The details of implementing a whitening filter depend a great deal on the specific observation characteristics: the contiguity of the data set, the uniformity of the sampling, etc. All whitening filters represent an attempt to use "out of transit" points to predict the flux "in transit"; i.e., whitening filters presuppose a knowledge of the correlation structure of the observation noise. Here we will assume that the noise is white or has been whitened. Now if the noise is not Gaussian, this detector may not be optimal. However, well-sampled photometric observations are often moderately characterized as Gaussian once outliers caused by cosmic rays and poor observing conditions are removed. In any case, time domain matched filters or their equivalent are the dominant detection strategies employed in this area. Thus, it is fruitful to consider this model given its popularity. We will further assume that the data have been treated in such a way that the transit pulse shapes are well preserved or that the effects of the prewhitening filter on the shape of the "whitened" transit are known. The search for transits of a given star's light curve, then, consists of convolving the light curve with a sequence of model transit pulses (distorted in the case of a prewhitener that does not preserve transit shape) spaced by each trial orbital period. Equivalently, the light curve may be convolved with a single model transit pulse and then folded at each trial period. The resulting detection statistics are examined for large positive values, the location of which gives the orbital period and phase of candidate planets. Equation (1) provides the formulation for a simple matched filter:

$$l = \frac{\boldsymbol{b}\cdot\boldsymbol{s}}{\sigma\sqrt{\boldsymbol{s}^T\boldsymbol{s}}} = \frac{1}{\sigma}\,\boldsymbol{b}\cdot\hat{\boldsymbol{s}}\,, \tag{1}$$

where $\boldsymbol{b}$ is the data vector, $\boldsymbol{s}$ is the signal to be found, and σ is the standard deviation of the zero-mean, white Gaussian noise (WGN). Note that this is simply the length of the projection of the data vector along the direction of the signal vector. Under the null hypothesis (no transits), l is a zero-mean, unit-variance Gaussian random variable. Likewise, it can be shown under the alternative hypothesis of $\boldsymbol{s}$ being present that l is a unit-variance Gaussian random variable with a mean equal to $\sqrt{E_s}/\sigma$. Here $E_s = \sum_i s_i^2$ is called the energy of $\boldsymbol{s}$. For transits consisting of rectangular pulse trains, equation (1) collapses into the square root of the number of points in transit times the mean data value during transit divided by the standard deviation of the observation noise.

In applying the detection algorithm, one will in practice construct a rather large number of detection statistics in order to densely sample the region of the parameter space of interest. For example, suppose we have 6 weeks of data from a ground-based program at a resolution of 4 hr^{-1} and 12 hr of observations each night and search for transiting planets with periods between 2 and 7 days. The step size in phase should be about $\frac{1}{4}$ of a transit duration, or 45 minutes. The step size in trial period should be set so that the farthest transits from a fixed central one do not shift more than about $\frac{1}{2}$ a transit duration from those for the previous trial period. The outermost transit pulses shift by $\frac{1}{2}$ the number of periods multiplied by the change in period. The average step size in period for this case is (3 hr/2)/(6 weeks/2/4.5 days) = 19 minutes, giving $\sim$373 trial periods. The average number of tests at any period is 4.5 days/(3 hr/4) = 144 tests. Thus, there are roughly

53,000 test statistics required per star to retain good sensitivity to all possible period/phase combinations. For 5000 stars, then, there are $\sim 3 \times 10^8$ test statistics constructed. The tests for each star are not independent, however, as every trial period will test for a transit at a given point in time for some trial phase. Thus the set of detection statistics for such a search is highly correlated and possesses a complex web of correlations.

This is illustrated by the following example. Consider star Cyg 1433 from the NASA Ames Vulcan Survey. Vulcan 1433 is a binary consisting of two late-F dwarfs undergoing grazing eclipses (Caldwell, Borucki, & Lissauer 2000). This star exhibits a transit-like feature with a depth of 3.19%, a duration of 3.36 hr, and a photometric period of 1.957 days (the orbital period is twice this value). The folded light curve for this star is displayed in Figure 1*a*, with the phase normalized such that the "transits" occur at a normalized phase of 0.25. By conducting a search for planets with orbits between 1 and 7 days on a grid with 7.5 minute spacing, we test 885,504 different models against the light curve. Figure 1*b* shows the maximum detection statistic obtained for each period sought for 2.5 hr transits. The maximum statistic obtained is 27.7 σ at a period of 1.96 days. Strong peaks are observed at rational harmonics of the fundamental photometric period, and the curve is elevated well above that of the bottom curve in the figure, which is the result for Cyg 1433's light curve once the transits are removed from the data. The multiple peaks in the top curve, which might be confusing at first sight, actually provide confirmation that the signal being picked up is caused by a periodic set of pulses of comparable depth. For most of the searches discussed in the remainder of this paper, we set up a nonuniform grid with respect to orbital period based on the following criterion: the correlation coefficient between a test at a given phase and period and the highest correlated test at the next largest period is no less than 0.75. This dictates the step size in period for a given period and number of transits observed and yields a maximum reduction in apparent S/N of only 12.5%.

We define the quantity $\boldsymbol{l}_{\max}$ as the maximum detection statistic over all tests of a light curve:

$$\boldsymbol{l}_{\max} = \max_i \{l_i\} . \quad (2)$$

The complementary cumulative distribution function (CCDF), $\overline{F}_{l_{\max}}(x) = 1 - F_{l_{\max}}(x)$, of $\boldsymbol{l}_{\max}$ interests us here.[2] The term $\overline{F}_{l_{\max}}(x)$ is the false alarm rate of a single search as a function of the detection threshold, x. The question is, how many independent tests, $N_{\rm EIT}$, were effectively conducted in performing the search? By this we mean, how many independent draws from an $N(0,1)$[3] process are required in order for the distribution of the maximum of the $N_{\rm EIT}$ draws to match the distribution of $\boldsymbol{l}_{\max}$ over some given range of the x-axis containing the desired false alarm rate? We call this process $\boldsymbol{N}_{\max}$ and the corresponding distribution $F_{N_{\max}}(x; N_{\rm EIT})$ and density $f_{N_{\max}}(x; N_{\rm EIT})$. We do not require that the two distributions match over the entire x-axis, just over the portion of interest.

The domain of interest warrants further discussion. The goal of this endeavor is to choose an appropriate threshold for individual tests. Strictly speaking, if the observation noise is WGN, the complementary distribution $\overline{F}_{l_{\max}}(x)$ provides this information directly; the value, x, of $\boldsymbol{l}_{\max}$ for which the sample CCDF $\overline{F}_{l_{\max}}(x) = N_{\rm FA}/N_{\rm stars}$ is the appropriate single-test threshold, where $N_{\rm FA}$ is the total number of false alarms. We note that in searching $N_{\rm stars}$ light curves for planets, we are performing $N_{\rm stars}$ independent searches. (If

[2] Throughout this paper the term density refers to the probability density function of a random variable. That is, given a random variable $\boldsymbol{y}$ (denoted by boldface italic type), the density or probability density function (PDF) is the function defined as the probability that an instance of $\boldsymbol{y}$ is confined to an infinitesimal interval about x: $f_y(x) = \lim_{\Delta x \to 0} \{P(x \le \boldsymbol{y} \le x + \Delta x)/\Delta x\}$. The term distribution refers to the cumulative probability distribution function (CDF), $F_y(x)$, where $F_y(x) = P(\boldsymbol{y} \le x)$. The term complementary cumulative distribution function (CCDF) refers to $\overline{F}_y = 1 - F_y(x)$.

[3] An $N(\mu, \sigma)$ distribution is defined as normal (i.e., Gaussian) with mean μ and variance σ^2.

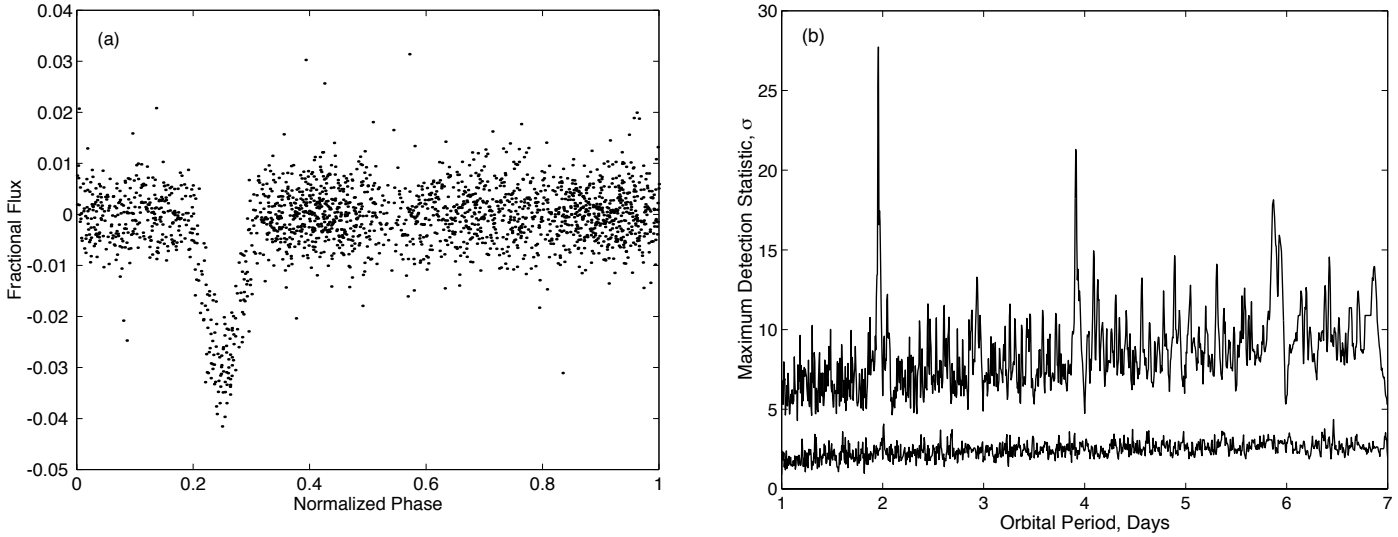

FIG. 1.—(*a*) Folded light curve for star Cygnus 1433 from the Vulcan campaign and (*b*) maximum detection statistics for a search for planets with orbits between 1 and 7 days for this star. The light curve is folded so that the transit-like feature occurs at a normalized phase of 0.25. The maximum detection statistic obtained for 2.5 hr transits is plotted for each period for the original light curve (*top curve*) and for the light curve obtained by removing the transits from the light curve (*bottom curve*). Note the sharp peaks appearing at multiples of the fundamental period of 1.96 days. The top curve is elevated above the bottom curve because there is some phase for each period sought corresponding to a model light curve with transits overlapping at least one of the features in the original light curve.

the searches are not independent, then something has gone wrong with the processing of the photometric data, as the resulting light curves should not be correlated, and hence, under the assumption that the observation noise is normal, the searches must be independent.) If we restrict the single-search false alarm rate to be $N_{\rm FA}/N_{\rm stars}$, the total expected false alarms is constrained to be equal to the desired $N_{\rm FA}$. This reasoning can be extended to individual tests as well. If the distribution of $\overline{F}_{l_{\max}}$ can be approximated by the distribution $\overline{F}_{N_{\max}}(x;\, N_{\rm EIT})$ in the region near $N_{\rm FA}/N_{\rm stars}$, then it is sufficient to choose the single-test false alarm rate to be $N_{\rm FA}/N_{\rm stars}/N_{\rm EIT}$ using the actual single-test statistics. Thus the region of interest is centered on $\overline{F}_{l_{\max}} = N_{\rm FA}/N_{\rm stars}$.

Now, to derive the distribution $\overline{F}_{N_{\max}}(x;\, N_{\rm EIT})$, we recall that the joint density of $N_{\rm EIT}$ independent Gaussian variables $X = \{x_i\}_{i=1,\ldots,N_{\rm EIT}}$ is

$$f(x_1, x_2, \ldots, x_{N_{\rm EIT}}) = \prod_{i=1}^{N_{\rm EIT}} g(x_i)\,, \tag{3}$$

where

$$g(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\,x^2\right) \tag{4}$$

is the PDF of an $N(0, 1)$ process (Papoulis 1984). The density of $N_{\max}$ can be obtained by noting that the probability of the maximum of $N_{\rm EIT}$ draws from an $N(0, 1)$ process attaining a value x is the probability of any one of the draws being equal to x times the probability that the remaining draws are less than or equal to x. As the draws are independent, we can write the density of $N_{\max}$ by inspection:

$$f_{N_{\max}}(x;\, N_{\rm EIT}) = N_{\rm EIT}\, g(x) G(x)^{N_{\rm EIT}-1}\,, \tag{5}$$

where

$$G(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\,y^2\right) dy \tag{6}$$

is the CDF of an $N(0, 1)$ process. The distribution of $N_{\max}$ is simply the distribution of an $N(0, 1)$ process raised to the $N_{\rm EIT}$ power:

$$F_{N_{\max}}(x) = G(x)^{N_{\rm EIT}}\,. \tag{7}$$

Thus, if the CCDF $\overline{F}_{l_{\max}}(x) = N_{\rm FA}/N_{\rm stars}$ at $x = \eta$,

$$N_{\rm EIT} \approx \log\left(1 - \frac{N_{\rm FA}}{N_{\rm stars}}\right) \Big/ \log\, G(\eta)\,. \tag{8}$$

If the joint distribution of the tests were known, the distribution of $l_{\max}$ could be found analytically or numerically, at least in principle. Given the correlation matrix $\mathbf{C}$ for the tests, the joint characteristic function is $\Phi(\Omega) = \exp(-\frac{1}{2}\Omega\mathbf{C}\Omega^t)$, but the joint density requires the inverse correlation matrix $\mathbf{C}^{-1}$ (Papoulis 1984). We note that the detection statistics are drawn from an $N_{\rm points}$-dimensional space, where $N_{\rm points}$ is the size of the data set. Hence, there can be no more than $N_{\rm points}$ linearly independent tests performed over the data set. However, the parameter $N_{\rm EIT}$ of the process, $N_{\max}$, may be much larger than the number of observations, $N_{\rm points}$, for a given sampling and planetary search, as will emerge from the examples considered later on. This underscores the fact that statistical independence of the tests conducted over a search space is separate from the linear independence of the signals considered as vectors in the underlying observation space. For the 6 week long observations considered above, there are only $\sim$2000 observations, with $\sim$53,000 tests applied to these points. Moreover, since there are more tests than points, $\mathbf{C}$ must be singular, and thus there does not appear to be a closed-form expression for the joint density of the tests. In any case, given the large size of the correlation matrix, integrating the joint density or joint characteristic function either analytically or numerically is impractical. Below we advocate the study of the distribution of $l_{\max}$ through Monte Carlo experiments.

Here we argue that the equivalent number of independent tests conducted per star, $N_{\rm EIT}$, is not determined by the distribution of the observation noise and is not strongly influenced by the presence of (red) colored noise. The Appendix contains a proof that the distribution of the observation noise does not affect the value of $N_{\rm EIT}$. Although the algorithm we provide in this section to estimate $N_{\rm EIT}$ is not affected by the actual noise distribution, the single-test threshold must be established by considering the actual distribution for the detection statistics, which is the subject of § 3 and § 4. We first note, however, that even if the observational noise is not Gaussian, we require that it be of bounded variance and that the light curves have been cleaned of strong, isolated outliers. Thus, the observation noise density should be well confined, even if the tails are longer than that for a Gaussian process with the same standard deviation. Second, we note that each detection statistic is a linear combination of several samples of observation noise. In most practical situations many samples "in transit" will be obtained simply by the fine sampling grid applied to ensure good sensitivity to the edges of transit events. For instance, the examples from the ground-based program we draw on feature sampling at $\geq 4\ {\rm hr}^{-1}$, giving at least eight points per transit for transits longer than 2 hr. Furthermore, we require in general that several (at least three) transits be observed. By the central limit theorem (Papoulis 1984), the density of the detection statistics may be well or moderately characterized as being Gaussian even in the event that the observation noise on individual data points is not. For example, let the observation noise $w(n)$ be white and drawn from the mixed Gaussian distribution with density

$$f(x) = \sqrt{\frac{5}{8}}\left[g\left(\sqrt{\frac{5}{2}}\,x\right) + \frac{1}{2}\,g\left(\sqrt{\frac{5}{8}}\,x\right)\right] \tag{9}$$

and with corresponding distribution

$$F(x) = \frac{1}{2}\left[G\left(\sqrt{\frac{5}{2}}\,x\right) + G\left(\sqrt{\frac{5}{8}}\,x\right)\right]. \tag{10}$$

In this case, $w(n)$ is a zero-mean, unit-variance process but is distinctly non-Gaussian. Now consider (1) single-transit statistics, l_1, for 3 hr transits (12-point pulses) and (2) three-transit statistics, l_3, for three 3 hr transits (36 samples from the mixed distribution). Figure 2 shows the CCDFs for $w(n)$, for l_1, for l_3, and for an $N(0, 1)$ process. Note that the single-transit and three-transit statistics are well modeled as being drawn from an $N(0, 1)$ process even though $w(n)$ is not an $N(0, 1)$ process.

For the case of red noise, if the correlation length of the noise were comparable to the length of a transit, we would expect $N_{\rm EIT}$ to be less than for the case of white noise. Consider a colored-noise process generated by passing a WGN process through a low-pass filter with impulse

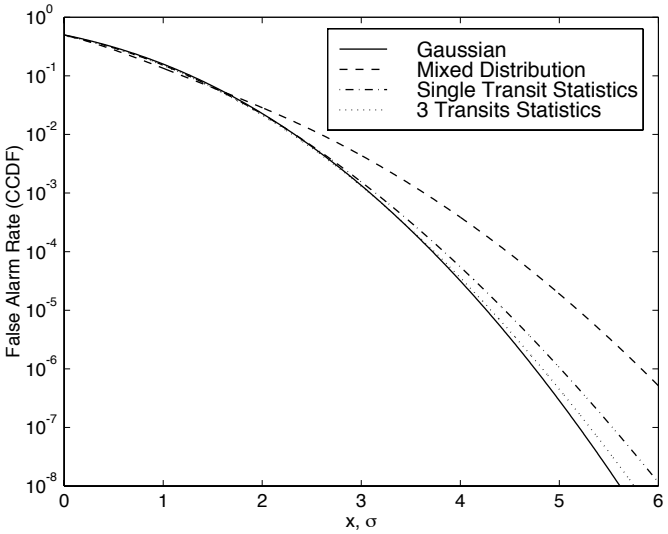

FIG. 2.—Sample and theoretical CCDFs (false alarm rates) as a function of threshold, x, for $N(0, 1)$ Gaussian noise (*solid line*); the mixed Gaussian distribution in the example in the text (*dashed line*); detection statistics for a single transit, l_1, in noise from the mixed distribution (*dash-dotted line*); and detection statistics for three transits, l_3, in noise from the mixed distribution (*dotted line*). The distributions from which l_1 and l_3 are drawn are 12-point and 36-point averages of samples from the mixed distribution, respectively. As more points in the mixed distribution are combined, the resulting distribution becomes more similar to a Gaussian one.

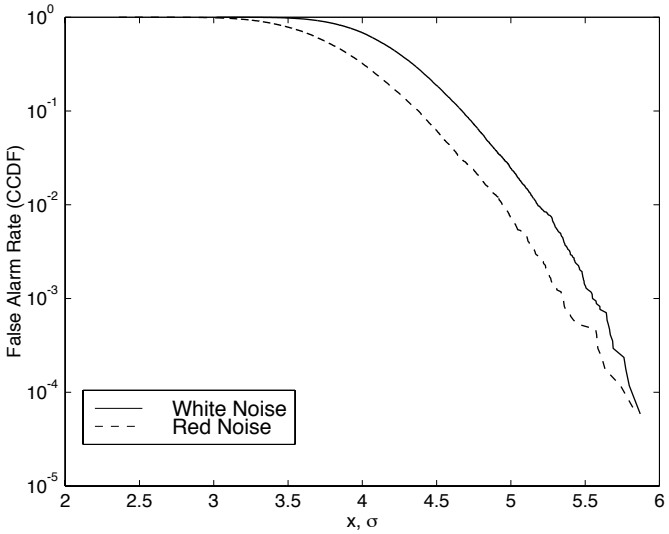

FIG. 3.—Sample CCDFs for search statistics for white and red Gaussian observational noise. The false alarm rate for red noise (*dashed line*) falls significantly faster than for white noise (*solid line*) as the threshold is increased. Thus, there are effectively fewer independent statistical tests conducted in searching the red-noise sequence for transits than there are in searching the WGN sequence.

response $\boldsymbol{h}$: $\boldsymbol{w}_c = \boldsymbol{w} * \boldsymbol{h}$. For this example, assume $\boldsymbol{h}$ is a rectangular pulse of length 3 hr. In applying a simple matched filter for single 3 hr transits, we convolve the observed noise, $\boldsymbol{w}_c$, with the unit-energy signal, $\hat{\boldsymbol{s}}$: $l_c(t) = k\boldsymbol{w}_c * \hat{\boldsymbol{s}} = k\boldsymbol{h} * \boldsymbol{w} * \hat{\boldsymbol{s}} = k\boldsymbol{h} * l(t)$, where $k = 1/\sqrt{E_{\hat{\boldsymbol{s}} * \boldsymbol{h}}}$ is a scale factor chosen to ensure that $l_c(t)$ is an $N(0, 1)$ process under the null hypothesis. The last term in the equality defining $l_c(t)$ shows that it is the (scaled) moving average of the single-event statistic $l(t)$ for white noise. The correlation length of $l(t)$ is half that for $l_c(t)$. Thus, as a time series, $l(t)$ has twice as many independent samples as does $l_c(t)$. Hence, we should anticipate that the expected maximum value for $l_c(t)$ is less than the expected maximum value for $l(t)$ for the same length observation. In fact, this should be true of any search for multiple transits as well, since multitransit statistics are linear combinations of single-transit statistics. This is borne out by a numerical example in which a 4 week observation is considered with a sampling rate of 4 hr^{-1} and a search for 3 hr transits with periods between 2 and 7 days is conducted. Figure 3 shows the CCDFs for both the red- and white-noise cases, demonstrating that the equivalent number of independent tests in conducting a full search is smaller for red colored noise than for white noise. That is not to say that it is easier to detect transits in colored noise. Although N_{EIT} is smaller, the scale factor k in effect reduces the S/N of a single transit by the same factor, making it more difficult to detect transits in colored noise with a correlation length comparable to a transit than it is for white noise.

As the assumption of white noise provides a conservative estimate for N_{max} in the case of red noise, let us consider WGN noise for the remainder of this section. Given the number of stars, N_{stars}, and the desired total number of false alarms, N_{FA}, we set the threshold so that the single-test false alarm rate is equal to $N_{\mathrm{FA}}/(N_{\mathrm{stars}}N_{\mathrm{EIT}})$. Let us consider some limiting cases for the complementary distribution of the maximum test statistic. Suppose there is a signal $\hat{\boldsymbol{s}}$ we test for in data set $\boldsymbol{b}$ such that $\boldsymbol{b} = A\hat{\boldsymbol{s}}$ and $\sigma = 1$. It follows that $l_{\mathrm{max}} = A^{-1}\boldsymbol{b} \cdot \boldsymbol{b} = A^{-1}\sqrt{\sum_i b^2{}_i}$. This will be the case, or nearly so, if we test for all possible signals or for a large number of signals that are dense on the N_{points}-dimensional unit hypersphere underlying the observations. Consequently, the distribution of $\boldsymbol{l}_{\mathrm{max}}$ would approach a χ-distribution with N_{points} degrees of freedom. This is the distribution for an incoherent matched filter or "energy" detector (Kay 1998) and explains its poor performance in comparison with a true matched filter. On the other hand, since the set of detection statistics for most transit searches is a complete set of vectors in the linear algebra sense, the distribution $\overline{F}_{l_{\mathrm{max}}}(x)$ is bounded below by $\overline{F}_{N_{\mathrm{max}}}(x; N_{\mathrm{points}})$. The search for planetary transit trains in most cases, however, is a rather restricted class of possible signals compared to the set of all possible signals. We should expect it to asymptotically approach the distribution for $\overline{F}_{N_{\mathrm{EIT}}}(x)$ for some $N_{\mathrm{EIT}} > N_{\mathrm{points}}$. While we do not supply a proof, we give several examples that demonstrate that $\boldsymbol{N}_{\mathrm{max}}$ does, indeed, provide a good model for the distribution of $\boldsymbol{l}_{\mathrm{max}}$ in the region of interest.

The algorithm for determining N_{EIT} is as follows:

1. For the distribution of observational time steps, construct a synthetic data sequence composed of independent identically distributed (i.i.d.) points drawn from a zero-mean unit-variance WGN process.

2. Examine the maximum detection statistic obtained from this synthetic data set by applying the simple matched filter algorithm of equation (1): $l_{\mathrm{max}} = \max_i \{\boldsymbol{x} \cdot \hat{\boldsymbol{s}}_i\}$ over the desired grid in the region of the period-phase duration parameter space of interest.

3. Repeat steps 1 and 2 a large number of times, at least several tens of the number of stars in the target sample.

4. Determine the number N_{EIT} of i.i.d. draws from a WGN process so that the complementary distribution of $\overline{F}_{N_{\mathrm{EIT}}}(x)$ matches the sample complementary distribution

function of the set $\{l_{\max}\}$ determined above at the point of interest $N_{\rm FA}/N_{\rm stars}$ (eq. [8]).

Note that it is not necessary to determine the value of $N_{\rm EIT}$ to exquisite precision as the CCDF of $N_{\max}$ falls rapidly at the false alarm rates of interest to transit photometry campaigns. Even relative uncertainties of 50% can be tolerated in the estimate of $N_{\rm EIT}$. The remainder of this section is devoted to several examples drawn from actual or anticipated observations.

2.1. *NASA Ames Vulcan Camera Observations*

We first consider the case of collecting data for a ground-based system similar to the NASA Ames Vulcan Camera where 12 hr of data are obtained per night at 4 $\rm hr^{-1}$ over several weeks. Figure 4 shows the results of conducting the Monte Carlo experiment above on 1, 3, and 6 week long sets of data, searching for planets with periods between 2 and 7 days. Over 10^5 trials were conducted for each data set. Taking $N_{\rm FA} = 1$ and a sample of 5000 stars, $N_{\rm EIT}$ is approximately 1900, 24,000, and 79,000, for 1, 3, and 6 weeks of data, respectively. Figure 5 shows how $N_{\rm EIT}$ evolves as a function of $\overline{F}_{l_{\max}}$ for each case. Although the search space is the same for all three data sets, the longer the baseline, the greater the resolution in terms of discriminating between planets with similar periods, and hence, the greater the number of effective independent statistical tests. That is, for longer data sets the correlation coefficient between one particular planetary signature and a second one drops off more rapidly as a function of period and phase as the parameters of the latter are varied from those of the former. Thus, the CCDFs, $\overline{F}_{l_{\max}}(x)$, for the 1 and 3 week long data sets "roll over" at smaller values of x than does the CCDF for the 6 week long data set. Not only are the values of $N_{\rm EIT}$ smaller for shorter data sets, the threshold required for the same false alarm rate is smaller as well. The sample CCDFs appear quite "ragged" at small values

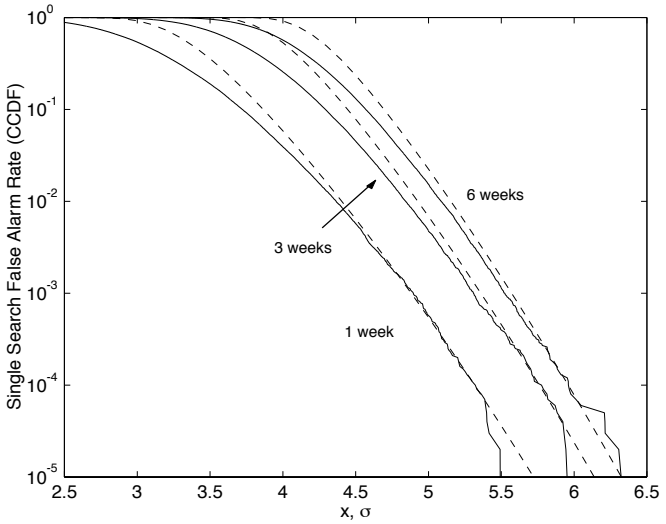

FIG. 4.—Sample CCDFs for planetary searches through 1, 3, and 6 weeks of Vulcan Camera data (*solid curves*) along with the theoretical curves for i.i.d. draws from a Gaussian process (*dashed curves*) that best match the empirical curves near a single-search false alarm rate of 1 in 5000. The effective number of independent tests performed in searching through data sets of these lengths, $N_{\rm EIT}$, is approximately 1900, 24,000, and 79,000, respectively.

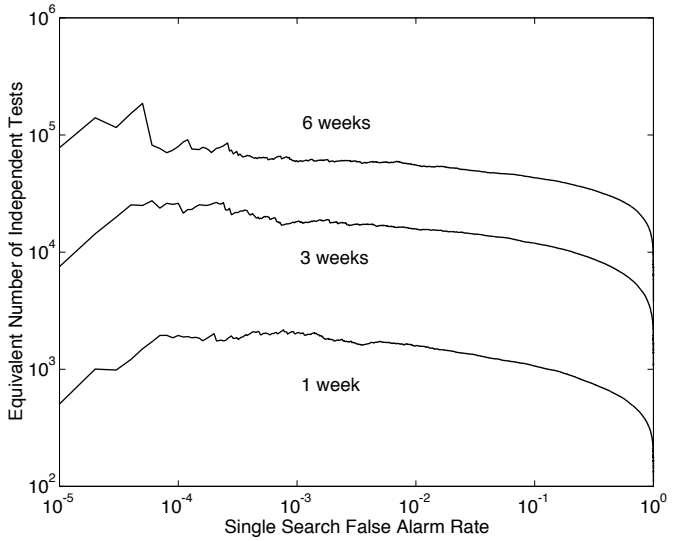

FIG. 5.—Equivalent number of independent tests for data similar to that collected by the Vulcan Camera as a function of the single-search false alarm rate for observations lasting 1, 3, and 6 weeks.

of $\overline{F}(x)$ because there are only a few samples available to estimate the behavior in the tail of the distribution. The number of trials performed to estimate the distribution must be high enough that a reliable estimate for $N_{\rm EIT}$ can be obtained at the relevant single-search false alarm rate.

2.2. *Multiple-Season Observations*

We next examine $N_{\rm EIT}$ for two 12 week seasons of Vulcan data and the *Hipparcos* data for HD 209458. The *Hipparcos* data consist of 89 points over 3 years' time, which is much sparser than the sampling for the Vulcan camera (>2000 points per season). Figure 6 illustrates the difference in the behavior of $\overline{F}_{l_{\max}}$ for each data set. Over 10^6 trials were performed in each analysis. The *Hipparcos* data are so sparse that in searching for planets with periods from 2 to 7 days, the sample complementary distribution $\overline{F}_{l_{\max}}$ is matched over a much shorter interval by $\overline{F}_{N_{\rm EIT}}$ compared to the two seasons of Vulcan data (Fig. 6*a*). This is illustrated in Figure 6*b*, where $N_{\rm EIT}$ is plotted versus the false alarm rate. At a single-search false alarm rate of 1/10,000, $N_{\rm EIT}$ is 110,000 for HD 209458 and is 790,000 for the Vulcan data. The *Hipparcos* data are so sparse that the signal space covered by the transit search is a significant fraction of the total surface of the 89-dimensional hypersphere underlying the signal vector space. Thus the CCDF rolls off much slower than that for $\overline{F}_{l_{\max}}$ until rather small false alarm rates are reached.

2.3. *The Proposed Kepler Mission*

The proposed Discovery-class *Kepler* mission would observe more than 100,000 target stars in the Cygnus constellation continuously for at least 4 yr at a sampling rate of 4 $\rm hr^{-1}$ (Borucki et al. 1997). The goal of the mission is to determine the frequency and orbital characteristics of planets as small as Earth transiting Sun-like stars. The range of periods of greatest interest is from a few months to 2 yr, with a range of transit durations from $\sim$5 to 16 hr for central transits of planets with periods over this orbital range. The average transit duration is 8 hr over these periods, assuming a uniform distribution of periods. (Note

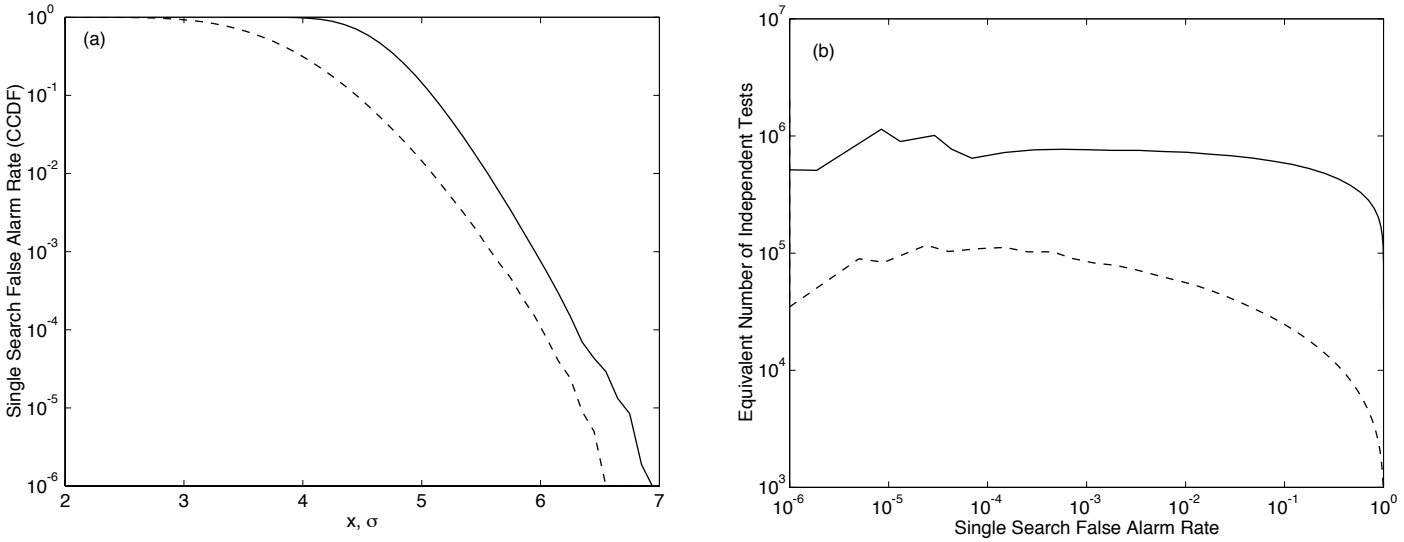

FIG. 6.—Analysis of multiple-year data sets. Panel *a* displays the CCDF's for the HD 209458 data set and for two 12 week observations with the Vulcan Camera spaced 1 yr apart. Although the Vulcan data consist of 4000 points, whereas the HD 209458 data consist of only 89 points, the effective number of independent statistical tests conducted in searching two seasons of Vulcan data set is only 8 times more than that for the HD 209458 data set, for a false alarm rate of 1 in 10^4. Panel *a* illustrates that the slope of the CCDF for the *Hipparcos* data set (*dashed curve*) is much different than that for the Vulcan data (*solid curve*). Panel *b* shows the evolution of $N_{\rm EIT}$ with false alarm rate corresponding to the Vulcan data (*solid curve*) and the *Hipparcos* data (*dashed curve*).

that since the average chord length of a circle of unit diameter is $\pi/4$, the average duration of a transit is $\pi/4$ times the duration of a central transit, which is 13 hr long at a period of 1 yr.) We applied the $N_{\rm EIT}$ algorithm to examine the statistics of $\boldsymbol{l}_{\rm max}$ for this experiment and to estimate $N_{\rm EIT}$. Figure 7 shows the result for over 10^6 searches for 8 hr transits for orbital periods between 90 days and 2 yr, yielding $N_{\rm EIT} \sim 1.7 \times 10^7$ for a single search. This agrees with the estimate obtained using K. Culler's approach discussed in § 1 and is no surprise as the assumptions for his method are met by this experiment. There is strong agreement between the theoretical curve and the empirical distribution of $\boldsymbol{l}_{\rm max}$, even for false alarm rates as high as 0.1. Thus, we estimate that there are $\sim 1.7 \times 10^{12}$ independent statistical tests required in performing the desired search over 100,000 stars. The corresponding requisite single-test threshold is 7.1 σ for no more than one expected false alarm for the entire campaign. The close agreement between the theoretical and the empirical curves most likely stems from the fact that the signals we are searching for are quite sparse on the unit-hypersphere underlying the 14,000-dimensional signal vector space for the simulations.

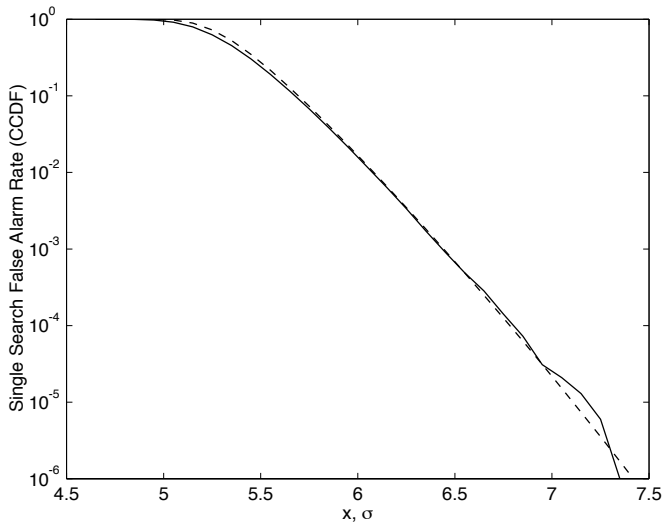

FIG. 7.—The sample CCDF for a 4 yr *Kepler* mission searching for 8 hr transits for planets with orbital periods between 90 days and 2 yr (*solid curve*), along with the theoretical curve for the maximum of 17 million draws from an $N(0, 1)$ process (*dashed curve*).

In the following sections we consider the problem of establishing an appropriate single-test threshold given the desired false alarm rate, the number of stars, and knowledge of $N_{\rm EIT}$.

3. BOOTSTRAP I: I.I.D. POINTS

Here we describe a technique for establishing the likelihood that a set of events was caused by chance where the noise is assumed to consist of i.i.d. points drawn from an unspecified noise distribution. This test is most appropriate when the photometric observations are well characterized as being white or are so sparsely sampled as to render detrending impractical. The photometry of HD 209458 in the *Hipparcos* catalog is an excellent example of the latter. The planet was detected initially by radial velocity variations, and ground-based photometry confirmed the planetary nature of these variations by measuring a planetary transit for the first time. Shortly thereafter, examination of the *Hipparcos* catalog revealed the presence of three transit signatures occurring at the correct times, leading to a much more accurate determination of the orbital period (Castellano et al. 2000). However, there were only 89 photometric measurements available for this star in the *Hipparcos* catalog over a 3 yr period. Further, the quality of the data set was marginal, as indicated by a standard deviation of greater than 1%. Figure 8 shows the *Hipparcos* data folded at the planetary period. The points in transit are not terribly conspicuous. Given that the transits should be about 0.6%–2.6% deep, and the numerous outliers in the data set, one might question the validity of the *Hipparcos* transits. Fortunately, there is a simple method for addressing the question of whether white i.i.d. noise generated the purported transits. Here is the *simple bootstrap algorithm*:

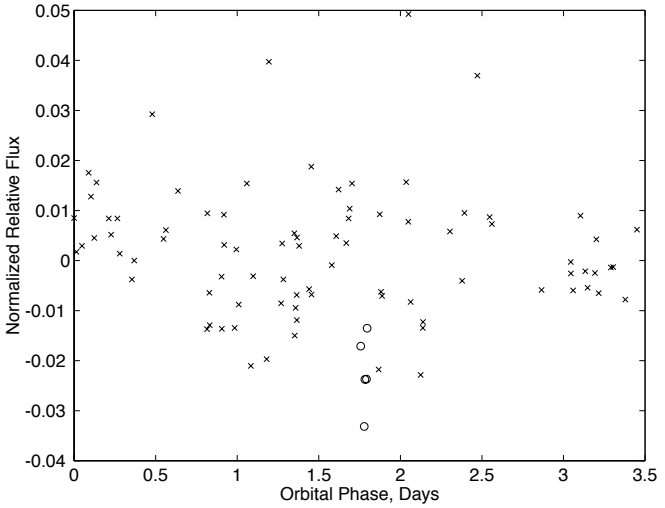

FIG. 8.—Folded light curve for HD 209458 from the *Hipparcos* catalog. The points in transit are denoted by the open circles.

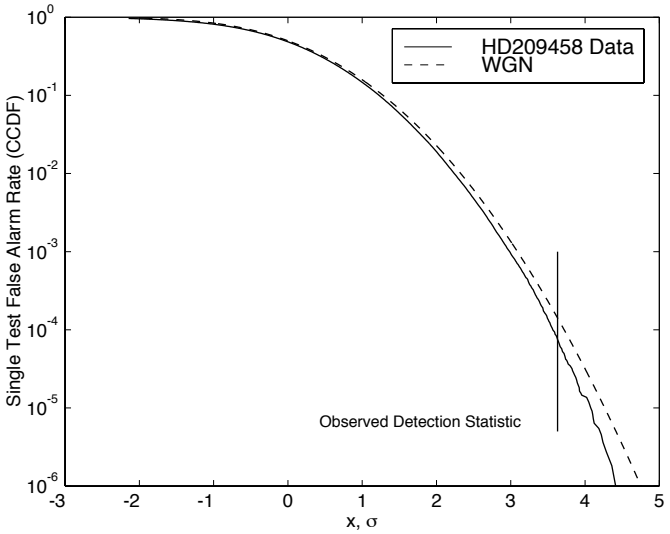

FIG. 9.—Single-test statistics for the simple bootstrap algorithm applied to the HD 209458 data set (*solid curve*). Note that the false alarm rate is bounded above by the distribution for $N(0, 1)$ noise (*dashed curve*). The observed detection statistic of 3.63 σ is denoted by the vertical solid line segment.

1. Draw k points at random with replacement from the data set, where k is the number of points in transit.
2. Form the detection statistic (eq. [1]) where the set $\{\boldsymbol{b}\}$ is the set of random points drawn in step 1.
3. Repeat steps 1 and 2 a sufficiently large number of times to address the question, at least several times the inverse of the desired false alarm rate. The fraction of times that the bootstrap statistics $\{l\}$ equal or exceed the observed detection statistic gives the confidence level for the detection.

For HD 209458, the detection statistic of the five transit points is 3.63 σ. (Here we normalize the detection statistics by sample standard deviation of the data set, rather than by the median absolute deviation as was done in Castellano et al. 2000. This does not alter the results in any way.) Running the bootstrap algorithm above yields approximately two trials in 100,000 that equal or exceed the observed statistic. Figure 9 shows the sample complementary distribution of the synthetic detection statistics along with that for $N(0, 1)$ noise. Note that the bootstrap method takes the shape of the transit events into account as well, although the results are not significantly altered if the shape of the transit is assumed to be a rectangular pulse. There is only one question for this data set, namely, are there transits at locations consistent with the later ground-based observations? The *Hipparcos* data set answers this question in the affirmative in no uncertain terms. One might consider, however, if there was a good chance of detecting the planet in the *Hipparcos* data in the absence of the a priori period and epoch. The method detailed in § 2 suggests that in searching for planets with periods between 2 and 7 days in the *Hipparcos* photometric catalog, one conducts $\sim$110,000 equivalent independent statistical tests for a single-search false alarm rate of 1/10,000. If the photometry for HD 209458 is typical of the quality of the photometric data in the *Hipparcos* catalog, one would expect almost all of the stars in the catalog to exhibit events with detection statistics as great as 3.63 σ. Another difficulty in searching such a sparsely sampled archive is that the results are ambiguous: there may be several combinations of phase and period that yield the same points in transit. Thus, the *Hipparcos* catalog does not represent a likely place to detect planets in the absence of a priori information. It may, however, provide fertile ground for future "precoveries," as in the case of HD 209458, or yield promising candidates for follow-up ground-based observations if judicious criteria are used to select a subcatalog from the full data set (Laughlin 2000).

4. BOOTSTRAP II: NON-I.I.D. POINTS

The bootstrap described in § 3 may not be the most appropriate test for establishing the significance of a set of candidate transits in a light curve. In most photometric data sets, especially those collected with ground-based telescopes, nature conspires to inject long-term variations into the data. Short-term fluctuations arising from photon counting statistics, cosmic-ray events, and atmospheric scintillation are well known and can be characterized fairly easily (Young et al. 1991; Dravins et al. 1998). The long-term variations, while well known, are not so easily characterized. The sources are transparency variations, seeing conditions changing over time, moving cloud structures in the field of view, changes in the performance of the instrument, intrinsic stellar variability, etc. The effect of these noise sources is to introduce correlations over timescales comparable to or longer than the duration of a transit and thereby invalidate the assumption of white noise. Such variations pose a major challenge to detecting planets with virtually all techniques, including photometry. In the context of radial velocity, Cumming et al. (1999) estimate that noise from intrinsic stellar variability is 2–3 times greater than the instrumental precision they achieved. Whereas the bootstrap in § 3 answers the valid question of how likely it is that the candidate transit series was caused by i.i.d. noise from the distribution underlying the observations, the presence of red noise (noise with a power spectrum inversely proportional to frequency) may invalidate the result. Luckily, the simple bootstrap can be modified to preserve long-term variations on timescales of interest to transit detection to answer a more appropriate question.

How likely is it that noise with a correlation structure and power similar to the observational noise would cause a detection statistic as high as that obtained for the candidate events?

In principle, this is not terribly different from the former bootstrap if we recall that searching for transits amounts to binning the data with a bin size equal to that of the trial transits where the centers of the bins move over small fractions of a bin. If the noise decorrelates over the timescale of a transit, then the two methods are equivalent for each individual test. The segmented bootstrap, then, consists of randomizing the occurrence of strings of data at least as long as the transits under scrutiny. Here is the *segmented bootstrap algorithm*:

1. Locate and remove (or fit and subtract) the candidate transits.
2. Remove all data strings shorter than a single transit (possibly all data collected on the nights of the candidate transits).
3. Pick a time tag in the remaining data set at random and collect the points within $\frac{1}{2}$ a transit duration of that point. This represents the first trial transit.
4. Repeat step 3 until the number of trial transits equals the number of candidates in the data set.
5. Form the detection statistic for the trial transits.
6. Repeat steps 3–5 a large number of times.
7. The significance of the candidate transits is equal to the fraction of times the detection statistic of a set of bootstrap "transits" equaled or exceeded the candidates' detection statistic.

We note that the algorithm can be accelerated tremendously by determining which points are within $\frac{1}{2}$ a transit of each time tag in advance and storing the sum of these points and the number of points in each set. In the case that the observational noise is not white and has not been whitened, the algorithm should be modified to include whatever detrending technique is applied to the light curves in the actual search. For comparison purposes with WGN, the statistics should be shifted and scaled so that they have a mean of zero and unit variance.

As an example of this technique we examine several light curves obtained from the NASA Ames Vulcan photometric search for extrasolar planets. The data were obtained on 59 separate nights spanning a period of 12 weeks. Figure 10 shows the observed light curves for seven stars in the Cygnus star field exhibiting transit-like features with depths of less than 3%. Note that Jovian-size planets transiting stars of spectral types from A0 to M0 produce changes in the relative flux from 0.2% for uninflated planets orbiting an A0 star to about 3% for an inflated planet orbiting a K5 star (Borucki et al. 2001). Thus the light curves shown in Figure 10 are readily mistaken with those for a transiting planet.

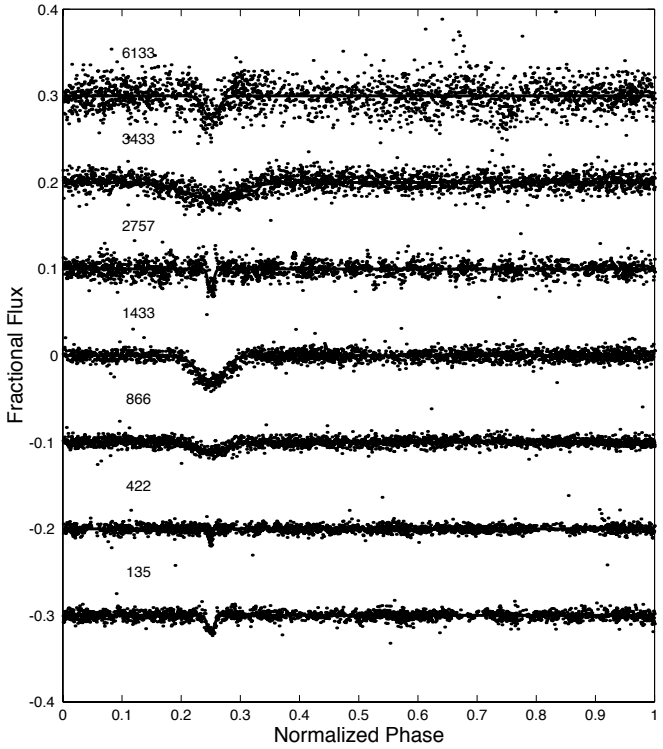

FIG. 10.—Folded light curves vs. normalized phase for seven stars in the Vulcan Cygnus field with transit-like features of 3% or less. The detected transit-like features occur at a normalized phase of 0.25. The scattered points are the measurements. The solid curves are the best-fit Gaussian pulses to the raw data points.

TABLE 1

VULCAN-CYGNUS STAR DESIGNATIONS

Vulcan Star	Tycho-2[a]	V[b]	$B-V$[b]	Spectral Type	Period (days)	Transit Depth (%)	max $\{l_{obs}\}$ (σ)	max $\{l_{boot}\}$ (σ)	False Alarm Probability
Cyg 135	2664-211-1	9.37	0.26	A3[c]	5.68	2.12	10.17	5.97	1.4E−24
Cyg 422	3141-2675- 1	9.90	1.17	K0[c]	6.08	1.27	8.17	6.01	1.6E−16
Cyg 866	2667-847-1	9.67	0.44	F5 V[d]	0.94	1.15	9.31	6.30	6.0E−21
Cyg 1433.....	2663-607-1	10.46	0.55	F9 V[e]	1.96	3.19	27.7	6.55	9.8E−169
Cyg 2757.....	2677-367-1	11.36	0.74	...	7.36	2.61	8.08	6.69	3.2E−16
Cyg 3433.....	3140-914-1	11.25	0.31	F0[f]	0.45	1.87	11.0	5.60	2.4E−28
Cyg 6133.....	2664-1385-1	11.62	0.76	G0[f]	2.58	2.56	8.66	7.83	2.3E−18

[a] Høg et al. 2000.
[b] Magnitudes and colors are from the *Tycho* catalog (ESA 1997), except for stars Cyg 2757 and Cyg 6133, the values for which are from the *Tycho-2* catalog.
[c] Cannon & Pickering 1923.
[d] Cyg 866 is a triple system of late-F dwarfs with two of the stars eclipsing and the third orbiting with an upper-limit period of ~340 yr (Posson-Brown et al. 2000).
[e] Cyg 1433 is a binary consisting of two late-F dwarfs undergoing grazing eclipses (Caldwell, Borucki, & Lissauer 2000).
[f] D. Latham 2000, private communication.

The star identifications, colors, and spectral types, where known, are given in Table 1. To determine if the masses of the companions suggested by the transit-like features are substellar, all of the stars except for Cyg 2757 are being observed spectroscopically by D. Latham and his colleagues at the Smithsonian Astrophysical Observatory (2000, private communication). Radial velocity results help rule out a planetary mass companion in all of the stars save one (Cyg 6133). Cyg 135 has a spectral type of A3 (Cannon & Pickering 1923) and is rapidly rotating, having a projected rotational velocity $v \sin i > 150$ km s^{-1} (D. Latham 2000, private communication). A mid-A dwarf has a radius of $2R_{\rm sun}$; therefore the observed 2% transit depth indicates a companion radius of at least 4 Jupiter radii. This size is not consistent with a planet-mass companion unless it is very young, $\sim 10^6$ yr (Guillot 1999; Burrows et al. 2000). Latham found that both Cyg 422 and Cyg 3433 are giants, which indicates that the observed brightness changes could not be the result of a transit of a planet-size companion. Stellar mass companions were observed spectroscopically around both Cyg 866, part of a triple system with one pair eclipsing (Posson-Brown et al. 2000), and Cyg 1433, an eclipsing binary. Both eclipsing pairs have orbital periods equal to twice the observed photometric period. Preliminary spectroscopic observations of Cyg 6133 indicate no radial velocity variation due to a stellar mass companion. The spectra indicate that the star is an early G slightly evolved off of the main sequence, with $\log g \sim 4.0$ (D. Latham 2000, private communication). Spectroscopic observations of this star are continuing; however, the photometric data suggest that this is an eclipsing binary. First, there is evidence for secondary eclipses in the light curve shown in Figure 10; second, the duration of the transit/eclipse is ~ 5 hr, twice what is expected for a planet orbiting a G dwarf. The last star, Cyg 2757, has not been observed spectroscopically because the flux variations are likely systematic noise, as no signal is seen at the expected period and phase in a second season of Vulcan observations. This star is presented as an example of where systematic noise can mimic a transit signal.

We applied the segmented bootstrap algorithm for transits of 2.5 hr duration to these stars to obtain the confidence level in the detections. Over 3.5×10^9 trials were performed for each star, and the resulting statistics were shifted and scaled to obtain zero-mean, unit-variance distributions. Table 1 contains the transit period, transit depth, maximum observed detection statistic max $\{l_{\rm obs}\}$ (shifted and scaled as were the null bootstrap statistics), the maximum bootstrap statistic max $\{l_{\rm boot}\}$ obtained for each star, and the probability of false alarm for the observed maximum statistic assuming the observation noise is WGN. All of the detections have high confidence, as none of the bootstrap statistics came within 0.83 σ of the observed statistic. Figure 11 shows the CCDFs for these seven stars along with that for $N(0, 1)$ noise. All but two of the curves cluster about that for $N(0, 1)$ noise. The curve for Cyg 2757 can be explained by noting that its light curve contains many outliers and highly correlated sequences of points that drop off from the average flux level at the edges of several nights' observations. The curve for Cyg 6133 can be explained in one of two ways: (1) If it is a grazing eclipsing binary, then the photometric period is the same as the dynamic period of the system and the secondary eclipses have not been removed prior to the bootstrap analysis. (2) If it is a high mass ratio system, then the photometric period is twice the dynamic period of the system and half the eclipses were not been discarded prior to generating the bootstrap statistics.

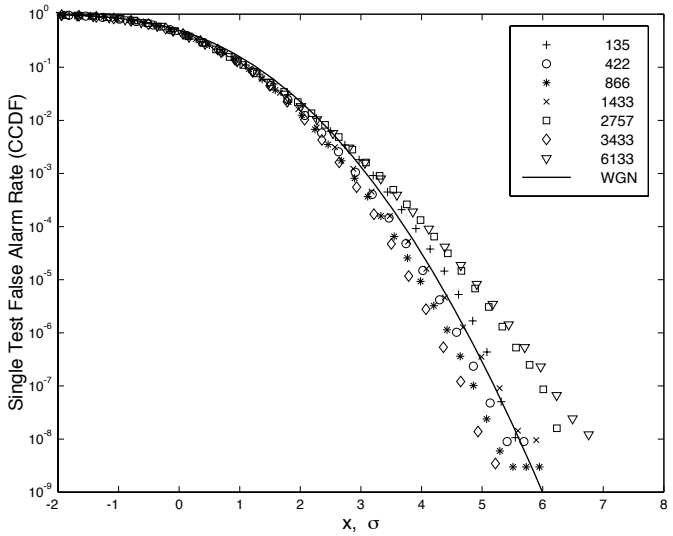

FIG. 11.—Bootstrap detection statistics for seven Vulcan target stars from the Cygnus field identified as possessing transit-like features (*various symbols*) and the expectation for unit-variance, zero-mean WGN (*solid curve*). All but two of the curves cluster about that for $N(0, 1)$ noise.

These two examples underscore the fact that the identification of candidates via a detection algorithm is not the end of the planetary detection process. Once identified, all candidates' light curves must then be fitted to analytic light curves to determine the uncertainty in the period, phase, and transit depth and to evaluate the likelihood that the source of the high detection statistic is indeed, a planet. As is the case for the Vulcan candidates, follow-up observations can provide valuable ancillary information that either provides direct confirmation of a planet or excludes the possibility of the source being a stellar mass object. Further photometric observations are also useful, as additional data allow for the identification of weak secondary eclipses and provide strong discriminatory power in cases where systematic effects are suspected. The segmented bootstrap analysis presented here also provides a method for identifying systematic errors, in that the empirical single-test CCDF should behave similar to that for $N(0, 1)$ WGN, otherwise there are statistically significant events occurring on timescales comparable to the transits being sought and at times not predicted by the most likely planetary candidate.

5. CONCLUSIONS

We have presented three Monte Carlo methods for estimating parameters that are essential to assessing the performance of photometric transit searches. The first method determines the equivalent number of independent statistical tests, $N_{\rm EIT}$, performed in searching a single light curve for transiting planets over a given range in orbital period for the desired total number of false alarms and the sample size, $N_{\rm stars}$. In establishing the value for $N_{\rm EIT}$, we used a Monte Carlo method which relied on random strings of $N(0, 1)$ noise. The value for $N_{\rm EIT}$ derived by the proposed algorithm does not depend on the actual noise distribution and is conservative for red noise. If the noise distribution is Gaussian, however, this procedure also yields the required single-

test threshold. As the single-test statistics for the real data sets examined here appear to be well modeled by the statistics for i.i.d. $N(0, 1)$ noise, this threshold is valid in cases even where the observation noise is not Gaussian. In the case of well-characterized non-Gaussian observation noise, the algorithm can be modified to construct synthetic light curves from the given distribution, rather than from an $N(0, 1)$ distribution, yielding the appropriate threshold. This may be difficult, however, as it is the distribution of the detection statistics that is required, not the distribution of the observation noise (see the Appendix). In the examples drawn from real observations presented here, this modification is not necessary.

The value for $N_{\rm EIT}$ can then be used in conjunction with the bootstrap methods of § 3 or § 4 to establish the appropriate single-test threshold for a photometric campaign. The product of $N_{\rm EIT}N_{\rm stars}$ gives the total number of independent tests performed in searching the entire data set for planets. If no further observations are possible, the value of the threshold should be chosen to constrain the total number of false alarms to be a small fraction of the total number of detections. However, if further observations of candidate stars are possible, then the threshold can be lowered in order to increase the detection rate so long as the follow-up observations can handle the increased observational load. These bootstrap methods estimate the receiver operating curves (ROCs) for a matched filter algorithm that takes into account in the best way possible the statistics underlying the noise of the observations. Which method is more appropriate depends on the nature of the observational sampling, with the simple bootstrap preferred in the case of highly sparse observations whereas the segmented bootstrap is more appropriate for observations with tightly clustered observations or evenly and densely sampled observations. The derived ROCs along with $N_{\rm EIT}N_{\rm stars}$ provide an objective approach for assessing the significance of any candidate planet and for estimating the detection rate of a proposed transit photometry campaign.

We are grateful to Dr. Laurance Doyle and Dr. Kent Cullers of the SETI Institute and Dr. Hans Jorg-Deeg of the Insituto de Astrofísica de Andolucía for their careful readings and constructive comments. We also thank Dr. Ron Gilliland of the Space Telescope Science Institute for reviewing the manuscript and for his help in improving the quality of the final version. This work was partially supported by funding from Origins and Advanced Project Offices at NASA Headquarters and the Astrobiology Office at NASA Ames Research Center.

APPENDIX

PROOF THAT THE DISTRIBUTION OF OBSERVATION NOISE DOES NOT AFFECT THE VALUE OF $N_{\rm EIT}$

Here we sketch a proof that the value for $N_{\rm EIT}$ does not depend on the noise distribution assumed for the observations. As discussed in the text, the distribution of the individual detection statistics may well be Gaussian even if the observation noise is not. For the purposes of the proof we will assume that the detection statistic, $\boldsymbol{l}$, is a function of an $N(0, 1)$ process, $\boldsymbol{x}$. Moreover, let us restrict $\boldsymbol{l}$ to be a zero-mean, unit-variance random variable. Let $\boldsymbol{l} = h(\boldsymbol{x})$ establish the relationship between $\boldsymbol{x}$ and $\boldsymbol{l}$, and let $h(\boldsymbol{x})$ be strictly monotonic increasing: $h(x_1) < h(x_2)$ iff $x_1 < x_2$. This does not limit the variety of noise distributions that can be considered as a function h can always be found relating two given distributions (Papoulis 1984). An example is $h(\boldsymbol{x}) = \boldsymbol{x}^3/\sqrt{15}$, the corresponding density of which possesses extremely long tails in comparison with an $N(0, 1)$ process. Indeed, even the sum of 100 independent samples has significant tails compared to an $N(0, 1)$ process. Now, the properties of $h(x)$ imply that

$$F_l(y) = P\{\boldsymbol{l} \le y\} = P\{\boldsymbol{x} \le h^{-1}(y)\} = F_{\boldsymbol{x}}(h^{-1}(y)) \; . \tag{A1}$$

Thus, there is a clear functional relationship between the distribution of $\boldsymbol{x}$ and the distribution of $\boldsymbol{l}$. Additionally, this functional relationship carries over to the maximum detection statistic over a given search, $\boldsymbol{l}_{\rm max} = \max_i \{l_i\}$, and the maximum of the corresponding Gaussian deviates, $\boldsymbol{x}_{\rm max} = \max_i \{x_i\}$:$\boldsymbol{l}_{\rm max} = h(\boldsymbol{x}_{\rm max})$. Hence, $F_{l_{\rm max}}(y) = F_{x_{\rm max}}(h^{-1}(y))$. Let $N_{\rm EIT}$ be the effective number of independent tests performed in searching for transiting planets for the Gaussian detection statistics $\{x\}$:

$$\overline{F}_{\boldsymbol{x}_{\rm max}}(x) = N_{\rm FA}/N_{\rm stars} \approx 1 - F_{\boldsymbol{x}}(x)^{N_{\rm EIT}} \tag{A2}$$

near $x = x_0$. But $x = h^{-1}(y)$ for some real number, y. So

$$\overline{F}_{\boldsymbol{x}_{\rm max}}(x) = \overline{F}_{\boldsymbol{x}_{\rm max}}(h^{-1}(y)) = \overline{F}_{l_{\rm max}}(y) \; , \tag{A3}$$

$$F_{\boldsymbol{x}}(x)^{N_{\rm EIT}} = F_{\boldsymbol{x}}(h^{-1}(y))^{N_{\rm EIT}} = F_l(y)^{N_{\rm EIT}} \; . \tag{A4}$$

Thus, $\overline{F}_{l_{\rm max}}(y) \approx 1 - F_l(y)^{N_{\rm EIT}}$ near $y = h(x_0) = y_0$. Therefore, the distribution of $\boldsymbol{l}_{\rm max}$ can be approximated by the distribution obtained from the process of choosing the maximum of $N_{\rm EIT}$ draws from the distribution of $\boldsymbol{l}$ in the region of interest, i.e., near $\overline{F}_{l_{\rm max}} = N_{\rm FA}/N_{\rm stars}$, which is the desired result.

REFERENCES

Black, D. C., & Scargle, J. D. 1982, ApJ, 263, 854
Borucki, W. J., Caldwell, D. A., Koch, D. G., Webster, L. D., Jenkins, J. M., Ninkov, Z., & Showen, R. L. 2001, PASP, 113, 439
Borucki, W. J., Koch, D. G., Dunham, E. W., & Jenkins, J. M. 1997, in ASP Conf. Ser. 119, Planets beyond the Solar System and the Next Generation of Space Missions, ed. David Soderblom (San Francisco: ASP), 153
Borucki, W. J., & Summers, A. L. 1984, Icarus, 58, 121
Brown, T. M., & Charbonneau, D. 2000, in ASP Conf. Ser. 219, Disks, Planetesimals, and Planets, ed. F. Garzón & T. J. Mahoney (San Francisco: ASP), 54
Brown, T. M., Charbonneau, D., Gilliland, R. L., Noyes, R. W., & Burroughs, A. 2001, ApJ, 552, 699
Burrows, A., Guillot, T., Hubbard, W. B., Marley, M. S., Saumon, D., Lunine, J. I., & Sudarsky, D. 2000, ApJ, 534, L97
Caldwell, D. A., Borucki, W. J., & Lissauer, J. J. 2000, in ASP Conf. Ser.

213, Bioastronomy 99: A New Era in the Search for Life in the Universe, ed. G. Lemarchand & K. Meech (San Francisco: ASP), 101
Cannon, A. J., & Pickering, E. C. 1923, Ann. Astron. Harvard College Obs. 98
Castellano, T., Jenkins, J. M., Trilling, D. E., Doyle, L. R., & Koch, D. G. 2000, ApJ, 532, L51
Charbonneau, D., Brown, T. M., Latham, D. W., & Mayor, M. 2000, ApJ, 529, L45
Cumming, A., Marcy, G. W., & Butler, R. P. 1999, ApJ, 526, 890
Deeg, H. J., et al. 1998, A&A, 338, 479
Deeg, H. J., Favata, F., & the Eddington Science Team. 2000, in ASP Conf. Ser. 219, Disks, Planetesimals, and Planets, ed. F. Garzón & T. J. Mahoney (San Francisco: ASP), 578
Doyle, L. R., et al. 2000, ApJ, 535, 338
Dravins, D., Lindegren, L., Mezey, E., & Young, A. T. 1998, PASP, 110, 610
ESA. 1997, The Hipparcos and Tycho Catalogues (ESA SP-1200; Noordwijk: ESA)
Gilliland, R. L., et al. 2000, ApJ, 545, L47
Guillot, T. 1999, Science, 286, 72
Henry, G. W., Marcy, G. W., Butler, R. P., & Vogt, S. S. 2000, ApJ, 529, L41
Høg, E., et al. 2000, A&A, 355, L27
Horne, J. H., & Baliunas, S. L. 1986, ApJ, 302, 757
Howell, S. B., Everett, M. E., Esquerdo, G. Davis, D. R., Weidenschilling, S., Van Lew, T., & Foxley, A. 2000, PASP, submitted
Jenkins, J. M., Doyle, L. R., & Cullers, K. 1996, Icarus, 119, 244
Jenkins, J. M., Doyle, L. R., & Deeg, H. J. 2000, Acta Astronautica, 46, 693
Kay, S. 1998, Fundamentals of Statistical Signal Processing: Detection Theory (Upper Saddle River: Prentice-Hall PTR)
Koch, D. G., Borucki, W., Webster, L., Dunham, E., Jenkins, J., Marriott, J., & Reitsema, H. J. 1998, Proc. SPIE, 3356, 599
Laughlin, G. 2000, ApJ, 545, 1064
Papoulis, A. 1984, Probability, Random Variables, and Stochastic Processes (New York: McGraw Hill)
Posson-Brown, J., Latham, D. W., Stefanik, R. P., Torres, G., Borucki, W. J., Caldwell, D. A., & Jenkins, J. M. 2000, AAS Meeting, 197, 4802
Robichon, N., & Arenou, F. 2000, A&A, 355, 295
Schneider, J., & Chevereton, M. 1990, A&A, 232, 251
Young, A. T., et al. 1991, PASP, 103, 221